Solution 1: Lasso Regularization

(a) First of all, we will use the fact that $\mathbf{X}$ has orthonormal columns to show that:

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \left(\underbrace{\mathbf{X}^T\mathbf{X}}_{\boldsymbol{I}} \right)^{-1} \mathbf{X}^T\mathbf{y} \\ \hat{\boldsymbol{\theta}} &= \mathbf{X}^T\mathbf{y} \end{aligned} \tag{1}$$

We will now use the result of eq. 1 to show that:

$$\begin{aligned} \arg\min_{\boldsymbol{\theta}} \mathcal{R}_{\text{reg}}(\boldsymbol{\theta}) &= \arg\min_{\boldsymbol{\theta}} \frac{1}{2} \sum_{i=1}^{n} \left(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{X}^{(i)} \right)^2 + \lambda \sum_{i=1}^{p} |\theta_i| \\ &= \arg\min_{\boldsymbol{\theta}} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \lambda \|\boldsymbol{\theta}\|_1 \\ &= \arg\min_{\boldsymbol{\theta}} \frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_1 \\ &= \arg\min_{\boldsymbol{\theta}} \frac{1}{2} \left(\underbrace{\mathbf{y}^T\mathbf{y}}_{indep.\ of\ \boldsymbol{\theta}} - 2 \underbrace{\mathbf{y}^T\mathbf{X}}_{\boldsymbol{\theta}^T} \boldsymbol{\theta} + \boldsymbol{\theta}^T \underbrace{\mathbf{X}^T\mathbf{X}}_{\boldsymbol{I}} \boldsymbol{\theta} \right) + \lambda \|\boldsymbol{\theta}\|_1 \\ &= \arg\min_{\boldsymbol{\theta}} -\hat{\boldsymbol{\theta}}^T \boldsymbol{\theta} + \frac{\boldsymbol{\theta}^T\boldsymbol{\theta}}{2} + \lambda \|\boldsymbol{\theta}\|_1 \\ &= \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{p} -\hat{\theta}_i \theta_i + \frac{\theta_i^2}{2} + \lambda |\theta_i|. \end{aligned} \tag{2}$$

(b) The advantage of this representation if we are interested in finding $\boldsymbol{\theta}$ is that we can optimize each $g_i(\theta_i)$ separately to get optimal entries for $\theta_1, \ldots, \theta_p$.

(c) Let's use the hint and compare $g_i(\theta_i)$ with $g_i(-\theta_i)$, for the case where $\theta_i > 0$:

$$\begin{aligned} g_i(\theta_i) &= -\hat{\theta}_i \theta_i + \frac{\theta_i^2}{2} + \lambda |\theta_i| \\ g_i(-\theta_i) &= +\hat{\theta}_i \theta_i + \frac{(-\theta_i)^2}{2} + \lambda |\theta_i| \end{aligned} \tag{3}$$

We also know that non-positive θ have always a greater or equal value for $g_i(\theta)$ than than their positive counterpart:

$$\theta_i > 0 \longrightarrow g_i(\theta_i) \leq g_i(-\theta_i) \tag{4}$$

The second and third term of both equations in eq. 3 are equivalent, so we can ignore them. Accordingly, to comply with the condition from eq. 4, the minimizer θ_i^* must be non-negative.

$$-\hat{\theta}_i \theta_i \leq \hat{\theta}_i \theta_i \longrightarrow \theta_i^* \geq 0 \tag{5}$$

(d) To calculate the minimizer θ_i^*, we will derive with respect to the parameter and set the derivative to zero:

$$\begin{aligned}\frac{\partial g_i(\theta_i)}{\partial \theta_i} &= -\hat{\theta}_i + \theta_i + \underbrace{\lambda}_{\theta_i > 0} \\ &= -\hat{\theta}_i + \theta_i + \lambda \overset{!}{=} 0 \longrightarrow \theta_i = \underbrace{\hat{\theta}_i - \lambda}_{\geq 0} \\ \theta_i^* &= \underbrace{\hat{\theta}_i - \lambda}_{\geq 0} \\ &= max(0, \hat{\theta}_i - \lambda)\end{aligned} \tag{6}$$

(e) Let's use the hint again and compare $g_i(\theta_i)$ with $g_i(-\theta_i)$, for the case where $\theta_i < 0$:

$$\begin{aligned}g_i(\theta_i) &= -\hat{\theta}_i\theta_i + \frac{\theta_i^2}{2} + \lambda|\theta_i| \\ g_i(-\theta_i) &= +\hat{\theta}_i\theta_i + \frac{(-\theta_i)^2}{2} + \lambda|\theta_i|\end{aligned} \tag{7}$$

$$\theta_i < 0 \longrightarrow g_i(\theta_i) \geq g_i(-\theta_i) \tag{8}$$

The second and third term in both equations of 7 are equivalent. Using the condition from eq. 8, we can conclude that the minimizer θ_i^* must be non-positive.

$$-\hat{\theta}_i\theta_i \geq \hat{\theta}_i\theta_i \longrightarrow \theta_i^* \leq 0 \tag{9}$$

(f) Calculating the derivative again and setting it to zero, we get:

$$\begin{aligned}\frac{\partial g_i(\theta_i)}{\partial \theta_i} &= -\hat{\theta}_i + \theta_i \underbrace{-\lambda}_{\theta_i < 0} \overset{!}{=} 0 \longrightarrow \theta_i = \underbrace{\hat{\theta}_i + \lambda}_{\leq 0} \\ \theta_i^* &= \underbrace{\hat{\theta}_i + \lambda}_{\leq 0} \\ &= min(0, \hat{\theta}_i + \lambda)\end{aligned} \tag{10}$$

(g)

$$\begin{cases}\theta_i > 0 \longrightarrow \theta_i^* = \overbrace{1}^{\text{sign}(\hat{\theta}_i)} \cdot max(0, \overbrace{\hat{\theta}_i}^{=|\hat{\theta}_i|} - \lambda) \\ \theta_i \leq 0 \longrightarrow \theta_i^* = min(0, \hat{\theta}_i + \lambda) = \underbrace{-1}_{\text{sign}(\hat{\theta}_i)} \cdot max(0, \overbrace{-\hat{\theta}_i}^{=|\hat{\theta}_i|} - \lambda)\end{cases} \tag{11}$$

By transforming the min function from 10 to a max function, we can combine the two cases in only one expression:

$$\theta_i^* = \text{sign}(\hat{\theta}_i) max(|\hat{\theta}_i| - \lambda, 0) \tag{12}$$